Clustering Assignment

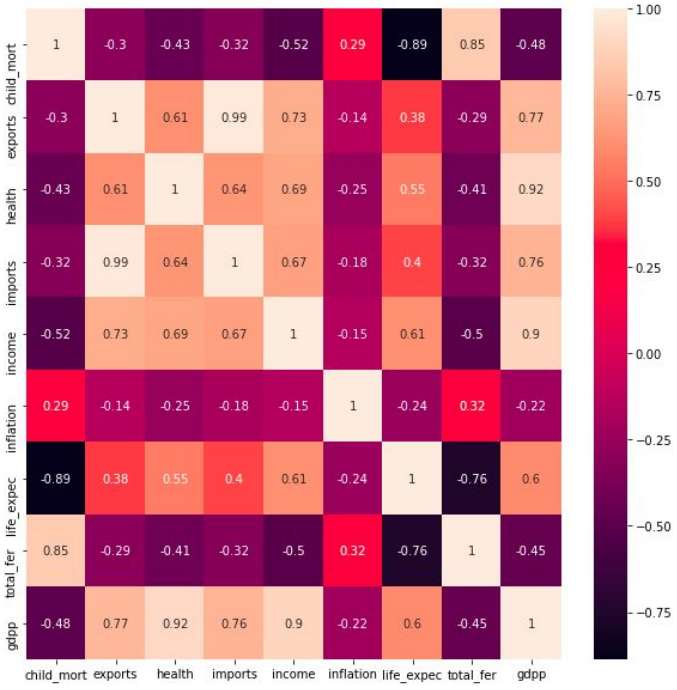
Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(what EDA you performed, which type of Clustering produced a better result and so on)

Sol: **Problem Statement**:

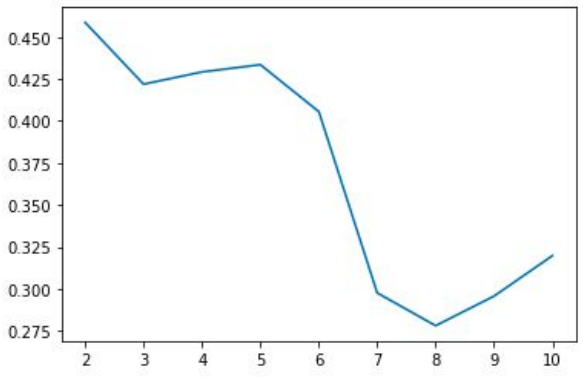
- The main aim of this case study is to categorise the countries using some socio-economic and health factors that determine the overall development of the country
- We need to find the countries by categorizing into clusters based on [gdpp, child_mort and income]

1. First we found the shape of the dataset and all the numerical columns in the dataset.
2. Then we performed Univatiate and Bivariate analysis.
3. Plotting the heatmap for all the numerical variables in the dataset gives the correlation variables int the data set.

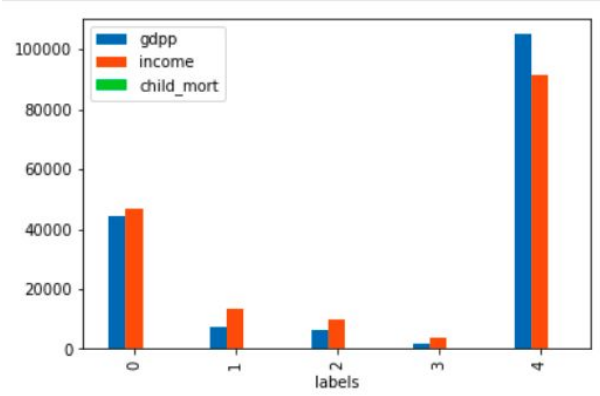


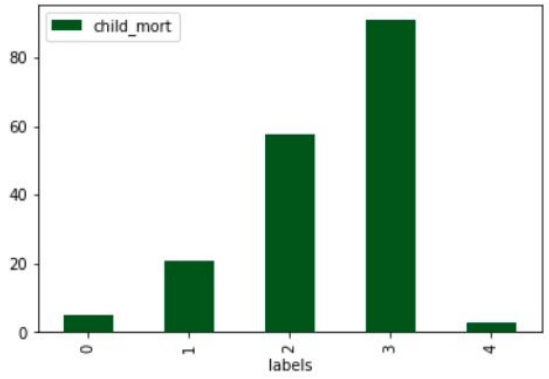
4. Inferences from above graph:
    a. We can see that health and gdp has highest correlation, followed by gdpp & exports.
    b. The least correlation is between life_expec & child_mort, followed by total_fer & life_expec

5. From the silhouette score graph we found that the highest value is for k=5 (ignoring k=2) followed by k=4 and k=3

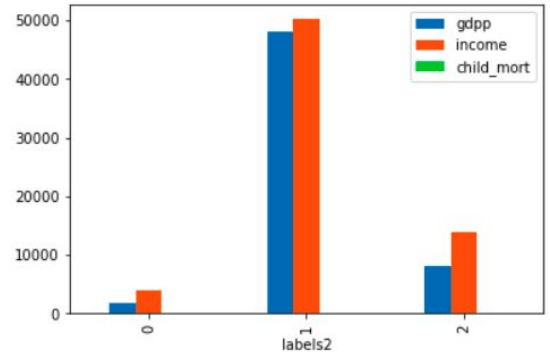


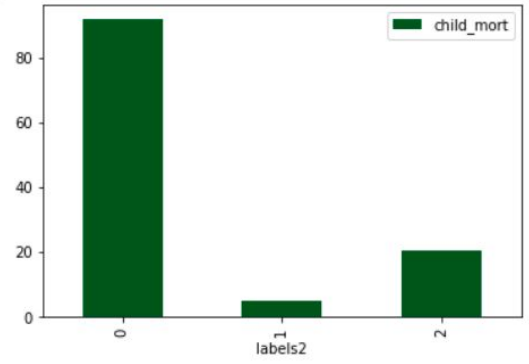
6. By performing profiling clustering for k=5





7. For k=3





8. Inferences:
    a. For the number of clusters=3, in cluster 0 we can see that child mortality is high, gdpp and income is very less
    b. So we can select the cluster 0 for our final reference
9. So by using hierarchical clustering we found that k=3 wil give the reasonable outputs for the dataset.
10. The top 5 countries that are dire need of help based on hign child_mort, low income, low gdpp are:

    a. Solomon Islands
    b. Iraq
    c. Botswana

d. South Africa
e. Eritrea

PART-B

1. **Compare and contrast K-means Clustering and Hierarchical Clustering.**

Sol: K-means Clustering:

- It is centroid based, partition-based method
- In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.
- One can use median or mean as a cluster centre to represent each cluster.
- Methods used are normally less computationally intensive and are suited with very large datasets.

Hierarchical Clustering:

- Hierarchical methods can be either divisive or agglomerative.
- In Hierarchical Clustering, results are reproducible in Hierarchical clustering
- Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
- Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.

2. **Briefly explain the steps of the K-means clustering algorithm.**

- Step **one**: Initialize cluster centers

  We randomly pick points and label them separately to represent the cluster centers.
- Step **two**: Assign observations to the closest cluster center

  Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center
- Step **three**: Revise cluster centers as mean of assigned observations

  Now we've assigned all the points based on which cluster center they were closest to. Next, we need to update the cluster centers based on the points assigned to them.
- Step four: Repeat step 2 and step 3 until convergence

  The last step of k-means is just to repeat the above two steps.

3. **How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.**

These methods are:

A. The Elbow Method
B. The Silhouette Method

A. The Elbow Method:

Within-Cluster-Sum of Squared Errors sounds a bit complex. Let's break it down:

- The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
- The WSS score is the sum of these Squared Errors for all the points.
- Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

Unfortunately, If we do not always have such clearly clustered data. This means that the elbow may not be clear and sharp.

So we'll go with Silhouette method

B. The Silhouette Method

$$s(\boldsymbol{o}) = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}$$

Where,

- ➔ **s(o)** is the silhouette coefficient of the data point o
- ➔ **a(o)** is the average distance between o and all the other data points in the cluster to which o belongs
- ➔ **b(o)** is the minimum average distance from o to all clusters to which **o** does not belong

There are main points that we should remember during calculating silhouette coefficient .The value of the silhouette coefficient is between [-1, 1]. A score of 1 denotes the best meaning that the data point o is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.

4. Explain the necessity for scaling/standardisation before performing Clustering?

Sol: All such distance based algorithms are affected by the scale of the variables.

We do not want our algorithm to be affected by the magnitude of these variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale. One of the most common technique to do so is normalization where we calculate the mean and standard deviation of

the variable. Then for each observation, we subtract the mean and then divide by the standard deviation of that variable:

$$z = \frac{x - \mu}{\sigma}$$

5. Explain the different linkages used in Hierarchical Clustering.
Sol: There are 3 types of linkages.

Single Linkage: For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

$$L(R, S) = min(D(i, j)), i \epsilon R, j \epsilon S$$

Complete Linkage: For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.

$$L(R, S) = max(D(i, j)), i \epsilon R, j \epsilon S$$

Average Linkage: For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.

$$L(R, S) = \frac{1}{n_R + n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i, j), i \epsilon R, j \epsilon S$$